Tier0/1 Status & Plans

2nd LCG Workshop

March 24th 2004

Tony.Cass@CERN.ch

Agenda

- Fabric Management
- Fabric Infrastructure
- Fabric Procurement

The LHC Computing Challenge

Summary of Computing Capacity Required for all LHC Experiments in 2007

source: CERN/LHCC/2001-004 - Report of the LHC Computing Review - 20 February 2001

(ATLAS with 270Hz trigger)

	CERN			Regional	Grand
	Tier 0	Tier 1	Total	Centres	Total
Processing (K SI95)	1,727	832	2,559	4,974	7,533
Disk (PB)	1.2	1.2	2.4	8.7	11.1
Magnetic tape (PB)	16.3	1.2	17.6	20.3	37.9

~6,000 PCs

Another ~1,500 boxes

c.f. ~2,200 PCs and ~350 disk servers at CERN today.

Only 1/3rd of the total capacity is at CERN… Grid Computing.

Fabric Management

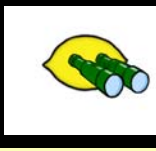
Fabric Management

- The box count brings difficulties for operating production quality systems
  - Tell us something new, I hear you say…
- At CERN, we break the problem down into three areas:
  - Installation and configuration—quattor (EDG/WP4)
  - System monitoring—Lemon (EDG/WP4)
  - Advanced hardware management—LEAF
- All interoperate as part of an overall Extremely Large Farm management system, ELFms.
  - e.g. quattor CDB populated automatically by the Hardware Management System part of LEAF.

quattor status & plans

- All core elements of quattor (CDB, NCM, SWREP, SPMA) were completed by late 2003.
- quattor has delivered measurable improvements for CPU fabric management at CERN.
  - LSF roll out; rapid security fix deployment.
- quattor is taking control of disk and tape servers
  - The "easy" 80% is complete; remaining 20% by Easter.
- quattor's CDB is now at the heart of asset management for the CERN Computer Centre.
- CERN will support quattor into LHC operation and help with deployment at other sites.
  - Can we collaborate on development of missing features?
    - » Security/authentification, graphical interfaces, …
      - (Beginners user guides!)

Lemon status & plans



- EDG/WP4 MSA and sensors deployed in 2002
  - Again, this delivered a significant improvement in the quality of service for our users.
- OraMon repository in production since July 2003.
  - APIs for data recording and data access.
- Focus now on delivery of monitoring data to the different interested parties:
  - Operators (alarm displays)
  - End-users
    - » Working with CMS in DC04 area
  - Hackers
    - » Via the repository API. Use this! Load nodes with processing, not monitoring!
  - Managers
    - » Me and Wolfgang…

LEAF status and plans

- Unlike quattor & Lemon, LEAF is CERN specific.
  - Remedy workflows adapted to our requirements.
    - » But the ideas are generic
- Hardware Management System
  - Boxes arrive: enter data in CDB, manage progression from installation through burn in tests to production.
    - » Integrated with our network database.
  - Manage hardware failure procedures & vendor calls.
- State Management System
  - "Give me N nodes. Make them like this. By then."
  - Initial implementation for kernel upgrades. More to come.
- Console management: CERN has adopted the SLAC console management software (which is also used by FNAL).
  - We do need some changes, but will work with Chuck Boeheim to integrate these into a common system.
  - A breach in the "Not Invented Here" wall?

Fabric Infrastructure

Fabric Infrastructure

Another 3 component problem!

- Space
- Power
- Cooling

Fabric Infrastructure

Or, rather (for CERN):

- **Cooling**
- **Power**
- **Space**

Sadly, the first two are closely linked…

Cooling

- The limiting factor on our ability to house systems.
  - 2MW in the 1500m² machine room is 1.3kW/m². Achieving more than this with air cooling is difficult.
  - Even this requires a conversion to underfloor air flow (which has just started), but if you have the space the flexibility gives the edge over water cooling.

Power

- What is the future for processor power? Inspite of reported "power budgets" per processor class, consumption seems to rise with each generation.
- CERN plans for 2.5MW active load; building consumption more like 5.5-6MW.
  - But with a 50% overcapacity in the low voltage distribution for flexibility.
  - Machine room & UPS consumption monitored by us (data stored in Lemon repository).
- Power factor as important as power.
  - Increased harmonics lead to unbalanced 3-phase system.
  - Fortunately EU directives seem to have led to an improvement from ~0.7 to ~0.9, even 0.95.
    - » We now reserve space for filters but don't include these in the baseline solution.

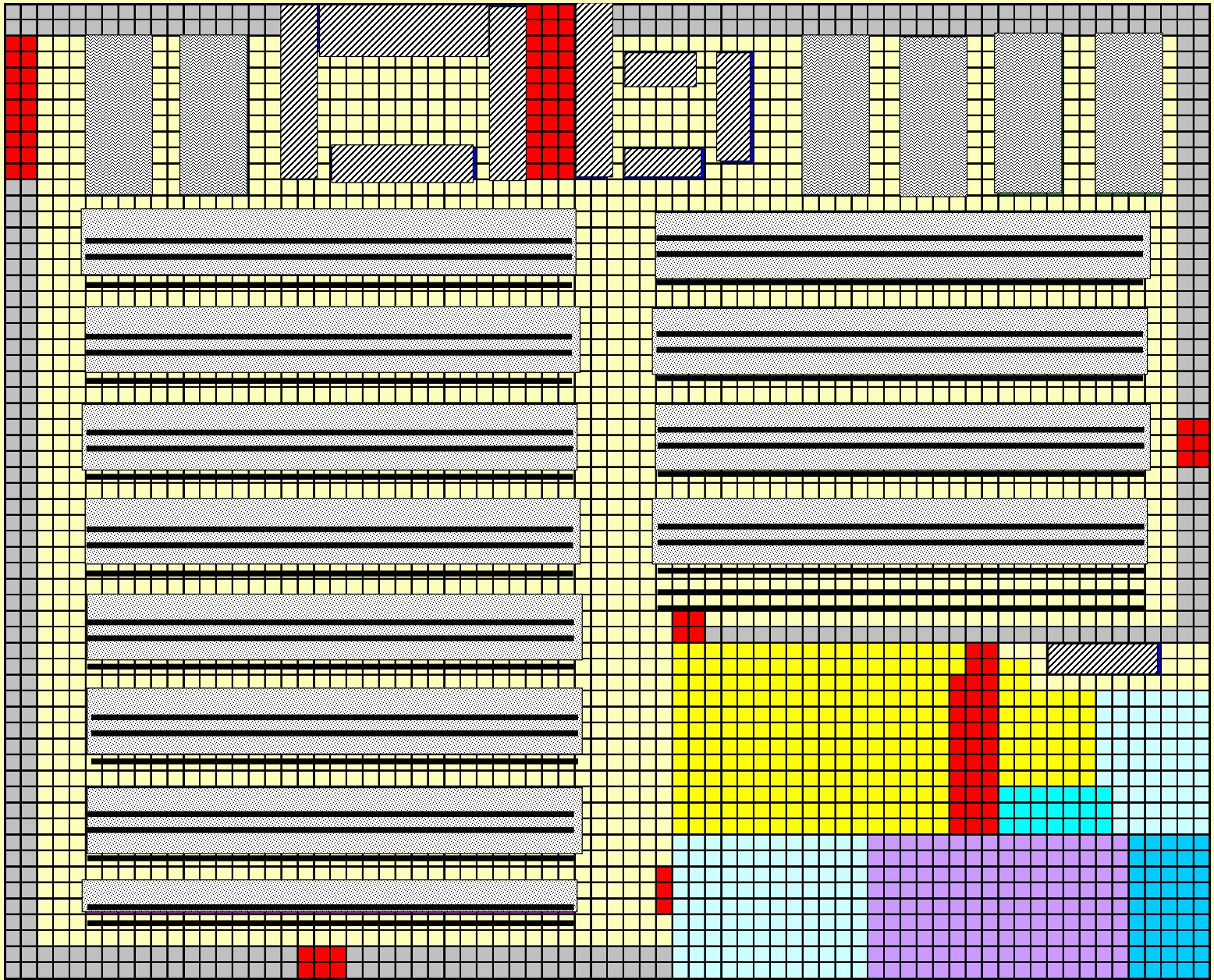








Space



Fabric Procurement

Fabric Procurement

- Procurement is a long process at CERN. To be able to install production systems early in 2006 we need to start the process now.
- But what do we want to buy?
- We assume same general fabric architecture as now: CPU, Disk and Tape servers interconnected with Ethernet hierarchy.
  - c.f. LCG milestones in June this year.
- Major concern at the moment is in the area of hardware quality.
  - Or, rather, the trade off between hardware quality and ongoing operation costs—Total Cost of Ownership…

TCO for CPU servers

- Which elements to include?
  - Reliability, Power, Space
- Space
  - CERN doesn't rent space, and we are generously equipped.
- Power
  - Blade systems offer some advantages, but they still have the same number of CPUs!
  - At current prices, consumption over 3 years only justifies a premium of 2-3CHF per Watt reduced consumption.
    - » Not a very green argument, though!
- Reliability
  - Hardware failures are ~10% of software rate.
  - What determines reliability? Can you prove systems are more reliable?
  - IT staff costs justify a premium of ~50CHF for systems with 0% hardware failure rate over 3 years.
- White box systems seem the cost minimum for us.

TCO for disk servers

- CERN has ~350 EIDE based disk servers for total capacity of ~250TB.
  - ☺ Cheap
  - ☹ Problem rate too high.
    - » Even discounting bad batch of Western Digital disks.
- But EIDE is dead anyway. How do we choose what we want to buy in 2006?
  - With confidence in the hardware quality!
- CERN has been testing SATA disks with CASPUR; can we profit from a wider collaboration?
  - But! We need hard evidence from large numbers of commercially purchased off the shelf arrays, not carefully selected individual systems.

Summary

Summary

- Fabric Management
  - On track. We are reaping the benefits of increased control through quattor and Lemon.
  - Small improvements (GUIs; access control) for these two areas; development is more concentrated now on advanced fabric management (LEAF).
- Fabric Infrastructure
  - On track. But how will CPU power consumption evolve?
- Fabric Procurement
  - 2006 is closer than you think! How slow is your bureaucracy?
  - What to buy?
    - » Hard not to see white box PCs as the cost optimum at CERN.
    - » Not at all so clear for disk servers. Advice (and collaboration!) welcome. Any volunteers?